NASA TECHNICAL MEMORANDUM

NASA TM 78149

(NASA-TM-78149) VECTOR STATISTICS OF LANDSAT IMAGERY (NASA) 19 p HC A02/MF A01 CSCL 14E

N78-15544

Unclas

G3/43 01845

VECTOR STATISTICS OF LANDSAT IMAGERY

By Robert R. Jayroe, Jr. and Debrah Underwood
Data Systems Laboratory

December 1977

NASA

George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
NASA TM 78149		

4. TITLE AND SUBTITLE

Vector Statistics of Landsat Imagery

5. REPORT DATE

December 1977

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)

Robert R. Jayroe, Jr. and Debrah Underwood

8. PERFORMING ORGANIZATION REPORT #

9. PERFORMING ORGANIZATION NAME AND ADDRESS

George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

10. WORK UNIT NO.

11. CONTRACT OR GRANT NO.

13. TYPE OF REPORT & PERIOD COVERED

Technical Memorandum

12. SPONSORING AGENCY NAME AND ADDRESS

National Aeronautics and Space Administration
Washington, D.C. 20546

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This work was sponsored by the Office of Applications' Data Management Program.
Prepared by Data Systems Laboratory, Science and Engineering

16. ABSTRACT

A digitized multispectral image, such as Landsat data, is composed of numerous four dimensional vectors, which quantitatively describe the ground scene from which the data are acquired. One of the purposes of this report is to investigate the statistics of unique vectors that occur in Landsat imagery and determine if that information can provide some guidance on reducing image processing costs. A second purpose of this report is to investigate how the vector statistics are changed by various types of image processing techniques and determine if that information can be useful in choosing one processing approach over another.

17. KEY WORDS

Image processing
Remote sensing

18. DISTRIBUTION STATEMENT

Unclassified — Unlimited

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
Unclassified	Unclassified	19	NTIS

TABLE OF CONTENTS

PRECEDING PAGE BLANK NOT FILMED

LIST OF ILLUSTRATIONS

LIST OF TABLES

VECTOR STATISTICS OF LANDSAT IMAGERY

I. INTRODUCTION

In one Landsat image there are 7 581 600 picture elements (pels), and each pel, which corresponds to a particular location of approximately 80 m resolution in the ground scene, is represented by a four-dimensional vector. The four components of this vector correspond to the reflected light intensity in each of the four different spectral images at a particular ground location. All four spectral images are digitized to 6 bits, but three of the images are radiometrically corrected to 7 bits. Thus, in three of the spectral images the data ranges from 0 to 127, while in the fourth image, the data ranges from 0 to 63. The maximum number of different four-dimensional vectors that could be generated with the previously mentioned combination of integers is $128^3 \times 64 =$ 131 096 512. However, since there are only 7 581 600 pels in an image, and some of the same vectors will occur many times, the total number of unique vectors will be less than 7 581 600.

The statistics that are examined are the number of unique vectors and the number of times that the unique vectors are repeated in a multispectral image as a function of ground scene, season, and test area size. These statistics are also reexamined after the original data are geographically corrected or compressed with various types of techniques.

II. TEST SITE DESCRIPTIONS

Three different test sites were examined. One test area was a Large Area Crop Inventory Experiment (LACIE) supersite in Finney County, Kansas, and can be described as an almost purely agricultural scene. The test area was 196 pels wide in the east/west direction and 117 pels long in the north/south direction, and contained a total of 22 932 pels. Eleven different passes of Landsat data were acquired over the test site during the period from October 22, 1975 to September 28, 1976.

Another test area was the Bald Knob, Tennessee Quadrangle, which could be described as a hilly rural area containing mostly agriculture and forest. This test site was 255 pels wide and 200 pels long, or a total of 51 000 pels, which is the approximate size of a 7.5 min quadrangle.

The third test area contained 1 440 000 pels, 1200 pels wide by 1200 pels long, and was bounded by the city of Mobile, Alabama in the north, Mobile Bay in the east, the Gulf of Mexico in the south and the Mississippi State line in the west. Six different passes of data were acquired during the period from October 17, 1972 to January 5, 1975. The October 17, 1972 pass is spotted with some clouds and haze, while the June 21, 1974 pass is spotted with fewer clouds. These two are the only images that contain any cloud cover. The ground scene contains a large variety of features (saltwater, freshwater, beaches, marshes, agriculture, forest, urban areas, etc.).

III. UNIQUE VECTOR STATISTICS

The unique vectors and their number of occurrences were extracted from the imagery using a program described in TM-78133 [1]. Table 1 gives the number of unique vectors and the pel to unique vector ratio (P/V) as a function of the number of pels for the test sites. This ratio can be used as a measure of image complexity, since for a given number of pels a more complex image would have more unique vectors.

A. Histogram Format

It is common practice to relate the processing costs of multispectral imagery directly to the number of spectral images and the number of pels. However, if it is recognized that what is actually being processed and interpreted is the unique vectors, then important processing cost reductions could be achieved if the processing costs were directly related to the number of unique vectors instead of the number of pels. The pel to vector ratio would be the factor by which the processing costs could be reduced.

One way to achieve the cost reductions is to use a histogram type format for the multispectral imagery. The histogram format consists of extracting all of the unique vectors and the number of times that they occur from a multispectral

TABLE 1. NUMBER OF UNIQUE VECTORS VERSUS NUMBER OF PELS

Number of pels	Number of Vectors	P/V	Date	Test Site
22 932	3 790	6.1	10/22/75	Kansas
22 932	3 304	6.9	11/8/75	Kansas
22 932	2 234	10.3	12/6/75	Kansas
22 932	1 995	11.5	1/1/76	Kansas
22 932	1 975	11.6	1/2/76	Kansas
22 932	2 254	10.2	2/6/76	Kansas
22 932	2 235	10.3	2/7/76	Kansas
22 932	6 576	3.5	4/18/76	Kansas
22 932	8 039	2.9	5/6/76	Kansas
22 932	6 213	3.7	6/16/76	Kansas
22 932	3 850	6.0	9/28/76	Kansas
51 000	11 179	4.6	4/14/73	Tennessee
1 440 000	63 688	22.6	10/17/72[a]	Alabama
1 440 000	31 751	45.4	11/17/73	Alabama
1 440 000	27 696	52.0	12/5/73	Alabama
1 440 000	75 331	19.1	4/10/74	Alabama
1 440 000	80 119	18.0	6/21/74[b]	Alabama
1 440 000	25 001	57.6	1/5/75	Alabama

a. Haze and spotted with clouds

b. Spotted with clouds.

image, or a portion thereof, and placing that information at the beginning of the data tape. The rest of the data tape is a description of the image that is accomplished by placing one number at each pel location which identifies the vector that belongs there. When multispectral image data are reformatted in this manner, it is not necessary to process a vector at every pel location. Instead, it is only necessary to process the unique vectors, and if image reconstruction is required, the results of processing each unique vector can be applied to every pel location using a table lookup procedure, which is very efficient. Table 1 indicates that the test sites could be processed from 3 to 58 times faster for classification inventories, density stretching, band ratioing, etc., if a histogram format is used.

B. Reduced Vector Representation

Table 1 also suggests that if significant cost reductions (by factors of hundreds or thousands) are to be achieved, then the number of unique vectors has to be reduced. This requires that a multispectral image be approximated with a reduced vector representation, and there are at least two reasons, in addition to the cost savings, for justifying this approximation.

First, it is observed that there will be tens or hundreds of thousands of unique vectors contained in a Landsat image, and the final desired product is usually a thematic map and inventory containing less than a hundred different classifications. Thus, the large number of unique vectors tends to represent an extreme overabundance of variations or information compared to the desired end result, and it is suspected that the same information could be extracted if groups of similar unique vectors could be approximated and replaced with an average.

Second, the statistics on the number of times that unique vectors are repeated in an image tend to support this approach. Regardless of the test site, there will be more unique vectors that occur once in the entire image than any other type. The next largest group is the unique vectors that are repeated twice, etc., and at the other extreme there will be many unique vectors that are repeated thousands of times in an image with each of these vectors having a different number of occurrences (i.e., there will be only one vector that occurs 2345 times, for example, only one that occurs 3013 times, etc.). Table 2 is the percentage of unique vectors that occur 15 times or less for the 19 images that were examined, and only the high and low values are shown. On a percentage

TABLE 2. PERCENTAGE OF VECTORS VERSUS NUMBER OF OCCURRENCES

Percentage of Vectors		Number of Occurrences	Minimum Accumulative Percentage
Minimum	Maximum		
33.43	49.41	1	33.43
12.35	17.94	2	45.78
6.79	10.16	3	52.57
4.33	6.43	4	56.90
3.08	5.15	5	59.98
2.41	3.46	6	62.39
1.82	3.00	7	64.21
1.38	2.28	8	65.59
1.02	2.71	9	66.61
0.90	2.33	10	67.51
0.68	1.42	11	68.19
0.54	1.32	12	68.73
0.34	1.11	13	69.07
0.44	1.27	14	69.51
0.24	1.02	15	69.75

image area basis, the vectors that occur a small number of times are relatively expensive to process, and, as a minimum estimate, 70 percent of the unique vectors would be expected to occur 15 times or less in the entire image. If, for example, a Landsat image could be satisfactorily approximated with 2000 unique vectors, then the processing costs could be reduced by a factor of approximately 3800. This approximation appears sensible when it is recognized that the original data and a reduced vector representation are, in reality, classification results with a relatively large number of classes, and that a final classification result with 30 to 40 classes is a reduced representation carried to an extreme. It is obvious that any image can be appropriately approximated to some degree without effecting the interpretative results. The main points of contention are the degree of approximation that is acceptable and how to perform the approximation.

C. Seasonal Dependence

There is one final observation concerning Table 1 that should be mentioned: the number of unique vectors exhibit a seasonal dependence. There are approximately three times as many vectors at the height of the growing season (spring) than there are in the nongrowing season (winter), and Figure 1 shows this dependence as a function of the months for the Kansas and Alabama test sites. Although the application of this information is not immediately obvious, it may be worthwhile to pursue, for example, on a per field basis to establish crop calendars where ground truth is not well known.

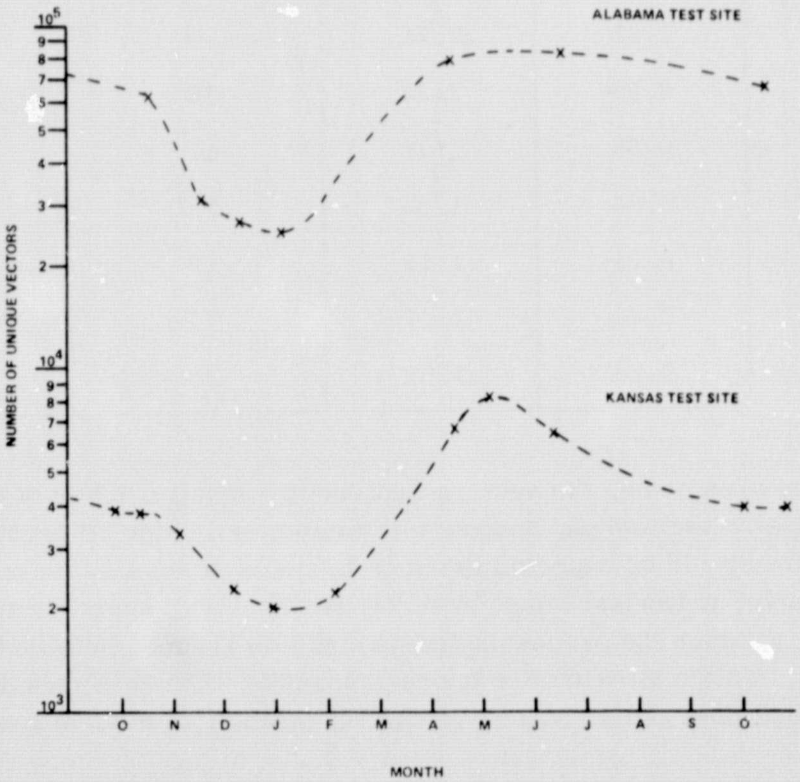

Figure 1. Number of unique vectors versus month.

IV. PROCESSING EFFECTS

A. Registration

The Tennessee test site was geographically corrected at approximately four times the resolution of the original data using the Nearest Neighbor (NN), Bilinear (BL) interpolation and Bicubic (BC) interpolation techniques. The correction techniques and other effects that they produce are described in TM X-73348 [2]. The resulting images were 560 pels wide and 555 pels long and contained a total of 310 800 pels. The corrected test site is square and rotated within the 560 × 555 pel image, and contains a total of 211 075 pels. The remaining pels are zero vectors used to make the resulting image rectangular, and they occur mostly at the corners. Figure 2 shows the number of unique vectors versus the number of pels for the three types of geographic correction techniques. The NN corrected test site contains the same number of unique vectors as the original data, and the interpolation techniques create new unique vectors as a result of spatial averaging. The graph also shows that as the extent of the spatial averaging increases (cubic versus linear interpolation), more unique vectors are generated.

Figure 3 shows the percentage of vectors that occur a given number of times for the original and geographically corrected data. The number of unique vectors that occur once in the bicubically corrected image (12 046) is larger than the total number of unique vectors (11 179) in the original image, while the number of unique vectors that occur once in the bilinearly corrected image is 7673. The graphs in Figure 3 are typical of what is obtained from most images, except for the NN graph. For the NN graph, the small percentages for one, two, and three occurrences are obtained from vectors at the edges of the test site. Since the resolution of the corrected data are approximately four times that of the original data, the NN graph will exhibit peaks at 4, 8, 12, 16, etc., number of occurrences due to repetition of vectors. In terms of image complexity, the geographic correction techniques that utilize interpolation create a corrected image that is more complex than the original data.

B. Compression

One of the images from the Kansas test site was compressed using several different compression techniques. The transform and difference methods techniques (Hadamard, H; Delta Pulse Code Modulation, DPCM; and Hadamard/Delta Pulse Code Modulation, H/DPCM combination) are described and discussed in

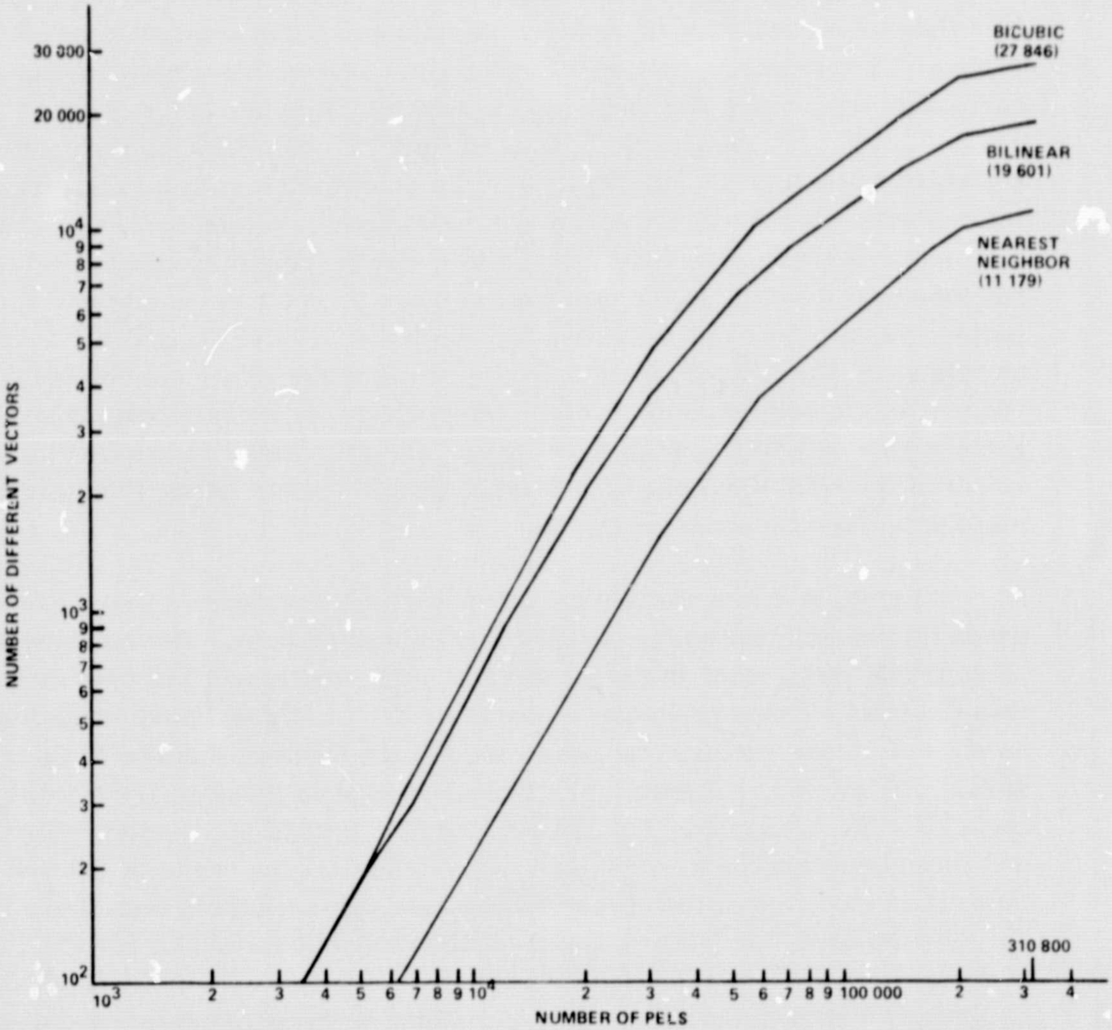

Figure 2. Registration technique versus image complexity.

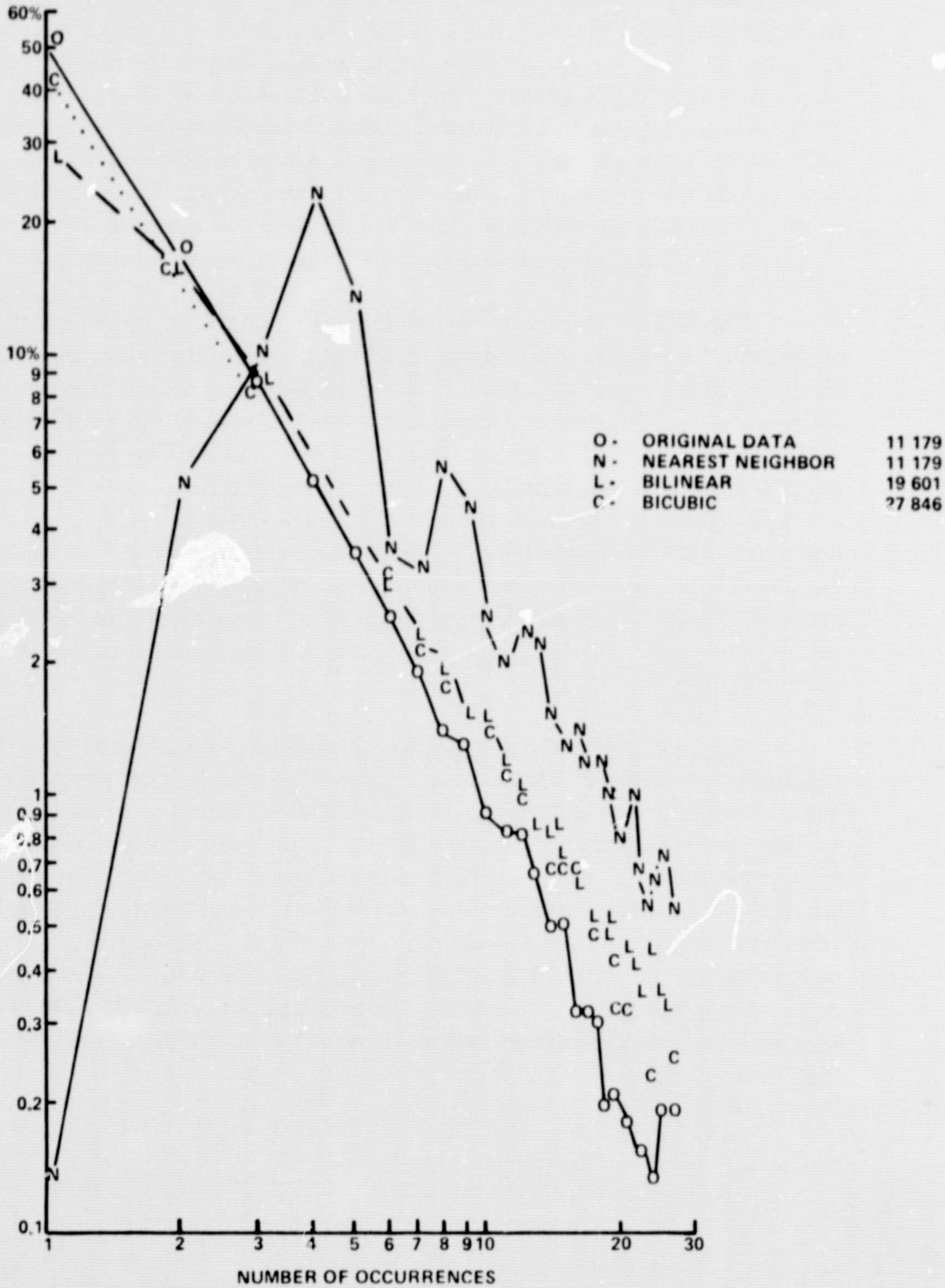

Figure 3. Percentage of different vectors having a given number of occurrences for registration techniques.

more detail in Reference 3, while the Cluster Coding Algorithm (CCA) is described in detail in TRW Final Report No. 26566. All of the approaches operated on each 16 × 16 pel array of an image, except for the CCA which can also use a 32 × 32 pel array. The basic difference between the two approaches is that the transform and difference methods approximate the distributional information extracted from the image data with a smaller number of bits and then reconstruct the image, whereas the clustering approach reduces the number of unique vectors in a 16 × 16 or 32 × 32 pel array to a specified number of average vectors which determines the number of bits required.

Figure 4 is a graph of the number of unique vectors versus the number of pels for the original and compressed data, which also show the number of bits used in the reconstruction of the image and the error that resulted from the approximation. The error (RMS) is the square root of the average of the variances for the four spectral images. The clustering approach is shown for cases where each pel array in the image is approximated with 8, 16, or 32 average vectors. The most obvious difference in the approaches is that the transform and difference method techniques create more unique vectors, while the clustering approach reduces the number of unique vectors. Thus, the process of approximating the distributional information extracted from an image with a fewer number of bits and reconstructing the image has the same effect as spatial averaging.

Figure 5 shows the percentage of vectors for a given number of occurrences in the original data and two compressed results that use a similar number of bits for image reconstruction and have almost identical errors. The DPCM method produces 4121 unique vectors that occur only once, which is more than the total number of unique vectors in the original data, while the clustering approach reduces the number of vectors occurring relatively few times. If a histogram format is used the cluster coded image could be processed five times faster than the original data and ten times faster than the DPCM reconstructed data. Thus, for the same amount of approximation error, the clustering approach produces an image that is ten times less complex than the DPCM approach.

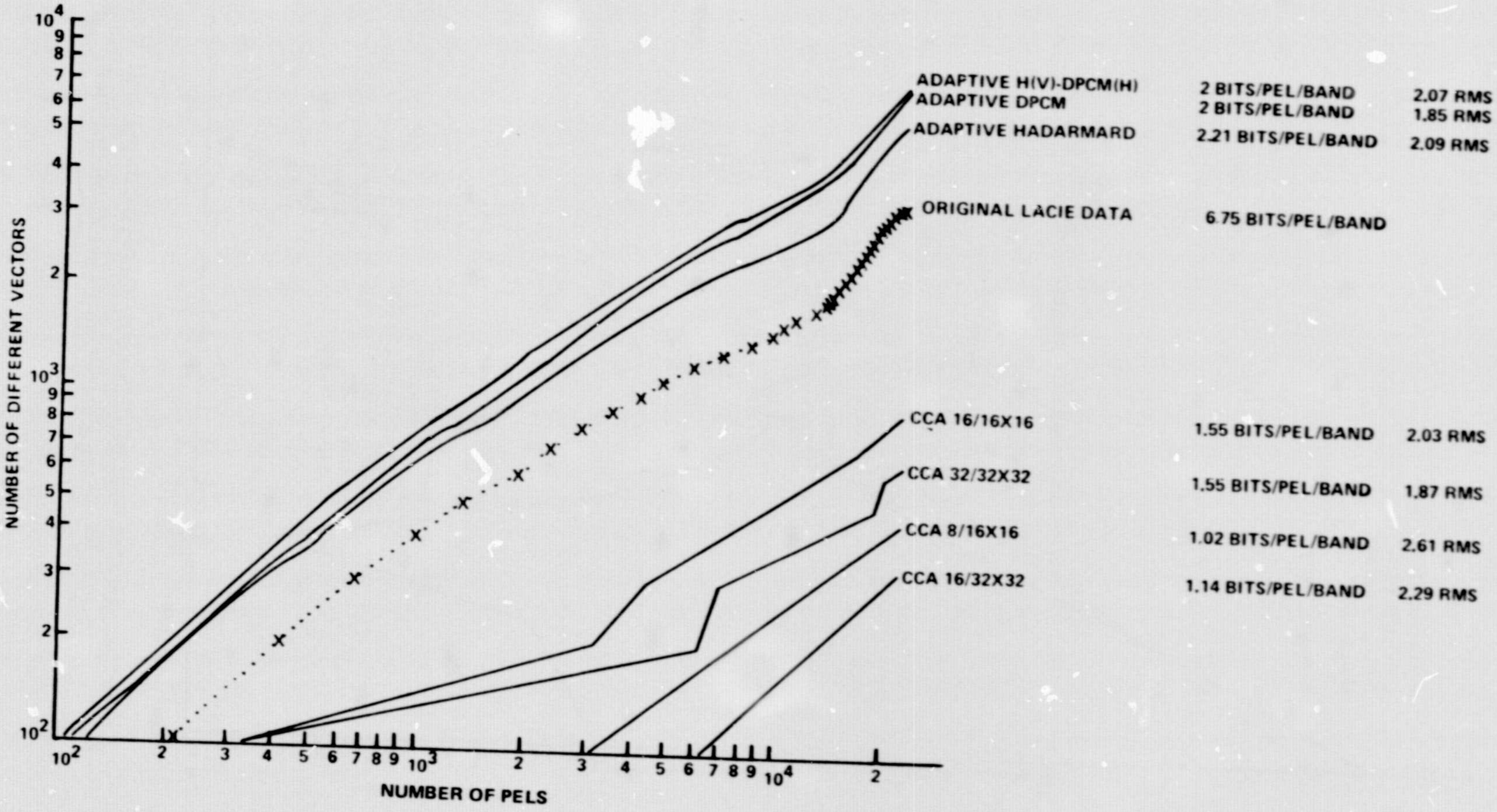

Figure 4. Compression technique versus image complexity.

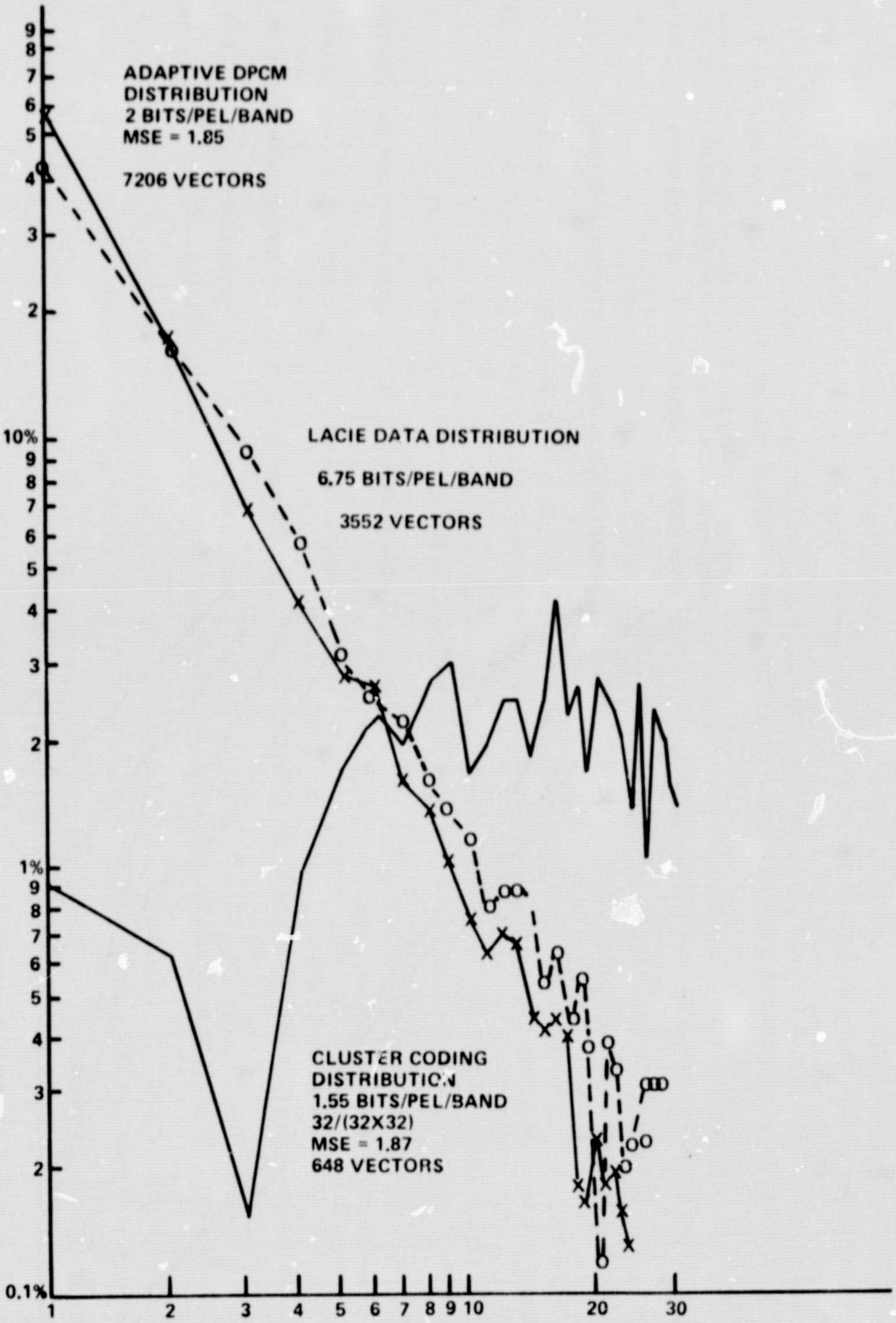

Figure 5. Percentage of different vectors having a given number of occurrences for compression techniques.

V. CONCLUSION

If large area resource inventories are to become a practical economic reality, there must be some mechanism to reduce the cost of image processing, especially for new sensors such as the thematic mapper which has two more spectral images and approximately four times as much data per image. In addition, there is already genuine concern that the cost of processing is prohibiting a majority of potential users from analyzing existing data and, therefore, considerably lessening its utility.

The use of a histogram format can lessen the cost impact by factors of ten in most cases without any information loss, but it also produces a constraint on the types of image processing that can be performed. Specifically, those processes that create new unique vectors via spatial averaging or an equivalent destroy the cost advantages of the format and, therefore, have to be eliminated from consideration.

If significant reductions (by factors of hundreds or thousands) are to be made in image processing costs, image data will have to be approximated by replacing the data with a reduced vector representation and using a histogram type format. Special purpose hardware devices can also be developed to reduce the processing costs even more. The main areas that need to be investigated are procedures for approximating image data and the degree of approximation that is acceptable.

REFERENCES

1. Jayroe, R. R.: A Fast Routine for Computing Multidimensional Histograms. NASA TM-78133, October 1977.

2. Jayroe, R. R.: Nearest Neighbor, Bilinear Interpolation and Bicubic Interpolation Geographic Correction Effects on Landsat Imagery. NASA TM X-73348, September 1976.

3. Study of Adaptive Methods for Data Compression of Multispectral Scanner Data. TRW Final Report No. 26566, NASA-Ames Contract NAS2-8394, March 1977.

4. Hilbert, E. E.: A Joint Clustering/Data Compression Concept. Ph.D. Dissertation, University of Southern California, May 1975.

APPROVAL

VECTOR STATISTICS OF LANDSAT IMAGERY

By Robert R. Jayroe, Jr. and Debrah Underwood

The information in this report has been reviewed for security classification. Review of any information concerning Department of Defense or nuclear energy activities or programs has been made by the MSFC Security Classification Officer. This report, in its entirety, has been determined to be unclassified.

This document has also been reviewed and approved for technical accuracy.

J. T. Powell

J. T. POWELL
Director, Data Systems Laboratory

U.S. GOVERNMENT PRINTING OFFICE 1977—740-193/194 REGION NO. 4